Journal of Physics: Conference Series

From Physics to industry: EOS outside HEP

To cite this article: X. Espinal and M. Lamanna 2017 *J. Phys.: Conf. Ser.* **898** 052023

From Physics to industry: EOS outside HEP

X. Espinal, M. Lamanna

CERN European Laboratory for Particle Physics 1211 Genève (Switzerland)

Abstract. In the competitive market for large-scale storage solutions the current main disk storage system at CERN EOS has been showing its excellence in the multi-Petabyte high-concurrency regime. It has also shown a disruptive potential in powering the service in providing sync and share capabilities and in supporting innovative analysis environments along the storage of LHC data. EOS has also generated interest as generic storage solution ranging from university systems to very large installations for non-HEP applications.

1. Introduction

While preserving EOS as an open software solution for our community we teamed up with COMTRADE company [1] within the CERN OpenLab [2] framework to productise this HEP oriented storage product in a fashion that promotes its adoption by other interested parties, notably outside HEP. In this paper we will deliver a status report of this collaboration and some example of the adoption of EOS outside HEP.

2. EOS: A Large Scale Storage System for the scientific community and beyond

EOS started its production phase in 2011 and currently holds 158PB of data and 1.1B files. It is a disk-only storage solution mainly focused on analysis and fast data processing with a very low access latency thanks to the multi-replication across nodes and JBOD layout [3]. Fast metadata access is guaranteed by the in-memory resident per instance namespace [4]. XROOT [3] is the principal access protocol while other protocols such as gridftp, http and fuse can be also enabled. Authentication is done by Kerberos/X509. The four big LHC experiments (ALICE, ATLAS, CMS and LHCb) are using intensively EOS and they do have a *private* instance while for non-LHC experiments they make use of a *shared* instance, where the main users (among others) are the experiments: AMS, COMPASS, NA48/62, nTOF, NA61 and CLIC. EOS is also the backend for the CERN file share and sync service: CERNBox [2]

[1] COMTRADE is a leading IT company with offices in 11 countries across Europe and North America. Comtrade is a system of companies that operate in the field of software solutions, system integrations and hardware distribution. With a proven track record of delivering industry-leading IT solutions and software engineering services, Comtrade has become a trusted developer of end-to-end technology products and solutions in various industries.

[2] CERN openlab is a unique public-private partnership that accelerates the development of cutting-edge solutions for the worldwide LHC community and wider scientific research. Through CERN openlab, CERN collaborates with leading ICT companies and research institutes

[3] The acronym stands for Just a Bunch Of Disks and can be seen as a collection of hard disks in a common enclosure that have not been configured to act as a redundant array of independent disks. The popularity of this layout has grown dramatically as raid systems performance are struggling with current commodity hard drive sizes [8]

[4] Currently evolving towards a distributed namespace infrastructure based on a shared K-V store system [1]

EOS has a reduced complexity with respect other large scale storage systems. This is mainly caused by its original design to have a very fast namespace without a complex database dependency and basing the structure on the existing and largely used within the physics community XROOT framework. The resultant product is a system which is not complex to administer, to maintain and even more important to scale. It can be adapted to several use cases and cater for different quality of services:

- Global capacity: the storage capacity of the nowadays commodity diskservers (many-TB hard drives and many-disk on the same enclosure) makes small storage set-ups very sensitive to hardware issues as they are constituted by few-servers. On the other hand if disk resources can be scattered among substantial number of nodes (≥ 10) the system will be running on a comfort zone in terms of replica distribution and fault tolerance. Summarising if the installation consist of a minimum number of storage nodes EOS can be used to accommodate from O(100TB) to O(100PB).

- Streaming capacity and concurrency: EOS file layout can be tuned to an arbitrary number of replicas per file hence accommodating to maximise the throughput needs and adapt to the concurrency requirements. This file layout can be set per directory hence being able to have a hot-data (multi-replica) and cold-data (few-replica) shared in the same storage system. Global throughput capacity increases in number of disks and number of replicas allowing easy scalability for multi-streaming.

- Diversity: No need for special hardware. No need for the same hardware. EOS can bind any kind of disk (commodity Hard Drives, SSD, Kinetic drives, etc.) as every disk is seen as a filesystem *per se* in the system. Also RAID setups are supported, either via RAID controllers or software RAID. As long as the storage device is seen as a filesystem by the system it can be attached to EOS.

- Reliability: Data loss in large scale storage systems is an unavoidable fact. Hardware failures, software bugs, silent corruption and operational (human) errors contribute to the data loss chain. These, together with today's data *explosion* leads to the massive storage installations being affected by the smallest failure probabilities. Hence design of storage systems need to take this into account. At CERN the Annual Failure Rate (AFR) observed during the last years for hard drives is approx. 2.5% and approx. 1% for raid controllers, both of which vary between different vendors (source [4] and [5, 6, 7]). On the disk side, double copy on different nodes with JBOD configuration systems like EOS are very resilient to hardware failures as the files are distributed across different nodes. On the other hand double copy single disk mirroring (RAID systems) configurations are heavily impacted by double disk failures [8].

- Money: Besides the pure storage capacity there is the need for few extra nodes to be set into the system. The headnode (metadata server) can be an standard node with an standard CPU and standard network interface. The only requirement is that it requires RAM memory to accommodate the namespace. Right new the average metadata footprints are 0.5kB per file and 1kB per directory (100M files and 1M directories it will required around 64GB to run smoothly). Ideally the headnodes should run in a pair to allow Master-Slave setup to increase fault tolerance and resilience.

3. EOS outside HEP: the idea, the project, the impact

Based on the experience we had at CERN since the start of the production phase of EOS, its continues growth and the adoption by all the experiments at CERN it was decided to start the *productisation* of EOS to make it suitable for non-HEP and eventually non-scientific environments. It was within this framework we started a project together with CERN OPENLAB and COMTRADE.

The main scope of the project is the evolution of the EOS system in the direction of a simplified use, installation and maintenance and to extend its utilisation by adding new supported platforms.
The initial phase that started on June 2015 put emphasis to provide a robust installation kit to allow rapid installation of EOS. The kit included the necessary installation, instructions and tools for operations such as administration guide and user guide. A test suite exercises the native EOS interface and the main access protocols: FUSE (Filesystem in User Space), Webdav/HTTP and S3).

The second phase which is starting now is focused mainly in a) integrating the new COMTRADE engineers into the development and operations team at CERN, gaining experience and autonomy on operations, maintenance and developing EOS to be able to provide first hand support, b) continue evolving the testing, installation and documentation to provide users the full fledge of EOS functionalities: Sync&Share(ie. CERNBox), erasure-coding and geographically distributed multi-site instances and c) provide a testing framework to run simultaneously after every release is built (nightlies+testing) to certify each EOS version as accurately and as quick as possible.

The impact of starting the productisation of EOS was immediate and several institutes within HEP and outside HEP manifested interest in adopting EOS as a storage solution.

- JRC [5]: The unit responsible for earth data observation studies decided to adopt EOS early 2016. The requirement came from the EU Copernicus Programme with the Sentinel fleet of satellites acting as a game changer for the data management and processing approach at JRC with an expected data rate of 10TB/day. Legacy systems were based on an appliance storage of 300TB accessed via NFSv4/CIFS. The current EOS set-up consists of 1.4 PB gross capacity, 10 storage nodes each with one JBOD of 24x6TB disks. System is configured to run in a double replica layout . The system will be extended to 6PB during the current year. More info of the EOS setup at JRC [9]

- Aarnet [6] The goal was to provide an scale-out filesystem underneath the ownCloud application [7], using the EOS fuse interface for file IO. The peculiarity and the challenge was that this installation is geo-distributed among Brisbane, Melbourne and Perth. The current EOS production implementation consists of two systems: a) **CloudStor** with 2.5PB presently and 12 machines with multiple MGM (metadata servers). The usage is cloud storage oriented and only 4% of the stored files are larger than 10 MB. And b) **Content Delivery Network** consisting of 300TBthat acts as a canary for CloudStor. This is extremely heavy reading with only one write client [10]

- IHEP [8]: the motivation to move to EOS was driven by several reasons: current storage systems manage metadata statically which leads to performance bottlenecks, as metadata and file operations are tightly coupled this makes it difficult to scale for a closed system. Traditional RAID technologies cause too much time consumption for data recovery and system crash in case of host failure. EOS has a very comprehensive file management capabilities: multiple copies, load balancing, etc. Current EOS implementations consists of three main clusters: **LHAASO** [9] : storage is used for LHAASO experiment batch computing. Current size is 230TB , 3.2M files and average size of 32MB. **Public EOS** Backend storage for IHEPBox based on Owncloud consisting of 145TB and holding 3.4M files with an average size of 4MB. Future plan is to extend EOS to serve as a storage

[5] European Commission Joint Research Centre https://ec.europa.eu/jrc/en/about/jrc-site/ispra

[6] Australia's Academic and Research Network https://www.aarnet.edu.au/

[7] ownCloud is an open source, self-hosted file sync and share app platform. Access and sync your files, contacts, calendars and bookmarks across your devices. https://owncloud.org

[8] Institute of High Energy Physics Chinese Academy of Science www.ihep.org/

[9] Large High Altitude Air Shower Observatory an experiment oriented to the study and observation of cosmic rays at Yangbajing, Tibet

system for other experiments at IHEP [11]

4. Summary

The scope of the project and its motivation to make EOS available outside CERN and outside HEP specific environments has been introduced. The prompt results resulting in various installations of EOS around the world in different areas supports our visions and manifests the *market* reality of Large Scale Storage Systems. EOS is being positioned as a valuable Storage technology ready to be adopted and with enough flexibility to cater for several Large Scale Storage needs.

References

[1] Peters A, Sindrilaru E and Adde " EOS as the present and future solution for data storage at CERN," J. Phys. Conf. Ser. **664** (2015) no.4, 042042. doi:10.1088/1742-6596/664/4/042042

[2] Mascetti L, Labrador H, Lamanna M, Moscicki J and Peters A"CERNBox + EOS: end-user storage for science," J. Phys. Conf. Ser. **664** (2015) no.6, 062037. doi:10.1088/1742-6596/664/6/062037

[3] "ROOT, an Object-Oriented Data Analysis Framework" http://root.cern.ch

[4] Numbers extracted from statistics collected at CERN CC. Private communications with CERN-IT Procurement Team members.

[5] dos Santos M, Waldron D "Observations made while running a multi-petabyte storage system" Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium

[6] Pinheiro E, Weber W and Barroso L "Failure Trends in Large Disk Drive Population" Proceedings of the 5th USENIX Conference on File and Storage Technologies, February 2007

[7] Schroeder B, Gibson G "Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?" Proceedings of the 5th USENIX Conference on File and Storage Technologies, February 2007

[8] Espinal X *et al.*, "Disk storage at CERN: Handling LHC data and beyond," J. Phys. Conf. Ser. **513** (2014) 042017. doi:10.1088/1742-6596/513/4/042017

[9] Burger A "EOS as storage back-end for Earth Observation data processing", EOS workshop 2-3 February 2017 CERN: https://indico.cern.ch/event/591485

[10] Jericho D "EOS at 6,500 kilometers wide", EOS workshop 2-3 February 2017 CERN: https://indico.cern.ch/event/591485

[11] Li H "EOS Usage at IHEP", EOS workshop 2-3 February 2017 CERN: https://indico.cern.ch/event/591485

[12] Peters A and Janyst L "Exabyte scale storage at CERN" J. Phys. Conf. Ser. **331** (2011) 052015. doi:10.1088/1742-6596/331/5/052015